CALCULATING WITH UNRELIABLE DATA IN BUSINESS ANALYTICS APPLICATIONS

Research in Progress

Beese, Jannis, SAP Research, St.Gallen, Switzerland, jannis.beese@sap.com

Bodner, Martin, SAP Research, St.Gallen, Switzerland, martin.bodner@sap.com

Abstract

The success of operational and managerial decisions depends on the reliability of the information provided to decision makers by the respective business analytics applications. Thus, in this research-in-progress paper, we explain how the mathematical foundations of the Algebra of Random Variables (AoRV) can be used to extend the capability of business analytics applications to process and report unreliable data. First, we present the theoretical foundations of the AoRV in a concise way that is tailored to business analytics. Second, we present and discuss two example cases, in which we evaluate an application of the AoRV to real-world business analytics scenarios. Initial results from this first design-and-evaluate feedback loop show that the additional reliability information provided by the AoRV is of high value for decision makers, since it allows to predict how uncertainties in complex business analytics scenarios will interact. As the next step of this research project, we plan to test the potential of the AoRV to extend business analytics applications through another evaluation loop in a fully natural setting.

Keywords: Business Analytics, Unreliability, Algebra of Random Variables, Design Science

1 Introduction

Business analytics systems provide analytical capabilities to manage and integrate different sources of business information and to analyze and report this data in a way that supports organizational decision-making (Davenport 2006; Elbashir et al. 2008). With the support of business analytics tools, management gets an aggregated glimpse of the operational processes and the financial health of the enterprise, thereby improving decision-making processes. However, the success and efficiency of managerial making decisions heavily depends on the reliability of the information provided to decision makers (Eisenhardt and Zbaracki 1992; Ford and Gioia 2000). We employ the term *data reliability*, to describe the extent to which a repeated, careful collection of data would lead to the same results (Murphy and Davidshofer 2001). Thus, while data reliability itself does not imply the validity of calculations (e.g., another issue is to consider all relevant data), data reliability issues limit the confidence in results (Murphy and Davidshofer 2001). Consequently, the added value that business analytics systems provide is dependent on the reliability of input data (Negash and Gray 2008). Especially in non-traditional controlling (e.g., related to innovation, employee motivation, or market perception), input data often stems from subjective estimates that are potentially biased (Molloy et al. 2011; Rindova et al. 2010). The data processed by corresponding business analytics systems may therefore contain non-negligible errors, resulting in a skewed depiction of reality (Parssian et al. 2004).

Since the use of business analytics systems is becoming increasingly pervasive across all management levels, data errors will likely incur a high cost for an organization (Redman 1998). Naturally, predictive models imply a certain degree of unreliability, which may be further compounded by mathematical aggregation processes (Draper 1995; Embrechts et al. 2013). Therefore, it seems evident to provide managers with additional information on the causes and the degree of data unreliability within predictive models to adequately manage unreliability in business analytics applications (Parssian et al. 2004; Wang et al. 1995; Wang and Strong 1996). One potential approach is to employ the mathematical foundations developed in the *Algebra of Random Variables* (AoRV) to calculate with unreliable data (Springer 1979). The AoRV provides a statistical foundation for performing mathematical operations on random variables, which may then be used to complement transformations and aggregations performed in business analytics applications. Employing the AoRV in the design of business analytics applications allows a mathematically precise treatment of data reliability aggregation, which would provide management with additional information on the reliability of key performance indicators. Consequently, the aim of this research project is to evaluate the applicability of AoRV to extend business analytics applications for an improved assessment of data reliability. We target the following research question:

How can the AoRV be applied to accurately assess data reliability in business analytics applications?

Methodologically, the research project follows the design science research (DSR) process of Peffers et al. (2007), in which we now report on the first evaluation loop (Abraham et al. 2014). Following Venable et al. (2016), we first employ a "quick & simple" (Venable et al. 2016, p. 82) DSR evaluation strategy, that is focused on assessing the applicability of the AoRV to real world business analytics scenarios. The next step will focus on evaluating the AoRV in a fully natural setting with real problems, real data, and real users (Venable et al. 2012).

In this research-in-progress paper at hand, we specifically report two initial contributions. First, we present the theoretical foundations of the AoRV in a concise way that is tailored to business analytics developers and IS scholars with a limited background in mathematics and statistics. Second, we discuss two typical cases, where we applied the AoRV to real-world business analytics scenarios by complementing extant data with additional reliability information to support strategic and operational decisions. The first case focuses on a typical strategic decision support scenario involving intangible assets related to human resource costs and benefit estimations. The second case evaluates the application of AoRV to predict product costs. For both cases, we evaluate the potential benefits of an application of AoRV in business analytics software practically through expert interviews and theoretically via a Monte Carlo simulation. Subsequently, we discuss both cases and the prerequisites and limitations of the AoRV approach.

Initial results show that the theoretical foundations provided by the AoRV can indeed be applied to solve common problems in real-world business analytics. For both cases, the additional reliability information is of high value for the company: calculations in the AoRV allow us to predict (i) how uncertainties in complex business analytics scenarios will interact, and (ii) whether they produce potentially dangerous compounding effects or whether different uncertainties tend to cancel each other out.

2 State of the Art and the Algebra of Random Variables

Nowadays, business analytics may best be described as a combination of techniques and tools that gather and analyze information to efficiently predict future outcomes (Bose 2009). Business analytics applications aim to give management a more effective way to interpret large datasets, by aggregating specific indicators to higher-level measures. Since the quality of available information has become of crucial importance for organizations (Wang et al. 1995), the reliability of data is discussed in many areas. Early works on business analytics systems (e.g., Zmud 1978) aim to empirically derive information quality dimensions and to develop organizing frameworks (e.g., DeLone and McLean 1992; Jarke and Vassiliou 1997). The dimensions evolved around four categories, comprising intrinsic, contextual, representational, and accessible quality (Lee et al. 2002). Other frequently used dimensions are accuracy, interpretability (Wang et al. 1995), completeness (Ballou and Pazer 1995), timeliness (Jarke et al. 1999; Lee et al. 2002), reliability, accessibility, and usability (DeLone and McLean 1992; Petter et al. 2013). There are also additional issues that arise when both, high and low quality data, are utilized to feed models that aggregate specific data inputs (Draper 1995). Consequently, the realized benefits of such business analytics tools and techniques directly depend on the reliability of the available input data.

Thus, scholars have developed various techniques to minimize data quality issues, such as data cleansing (Hernández and Stolfo 1998), data source calculus and algebra (Parssian et al. 2004), and dimensional gap analysis (Kahn et al. 2002). Similarly, organizations and managers have worked to establish "acceptance" levels with regard to their data (Madnick et al. 2009). Wang and Strong (1996) found that managers do not to consider reports provided by their business analytics systems, if they do not have trust in the reliability, accuracy, or origin of the data. The increasing pervasiveness and importance of business analytics applications in modern organizations has further raised these expectations of management with regard to accuracy and responsiveness (Shankaranarayan et al. 2003). Providing additional information about data quality to managers could increase the awareness of relational factors and uncertainty during the decision-making process in complex business analytics scenarios. Consequently, this research-in-progress report evaluates the application of AoRV to enrich business analytics applications with additional information about data reliability.

The AoRV provides a theoretical basis for this paper by defining a framework that allows mathematical manipulations of random variables. Similarly to how calculations with regular numbers work (e.g., adding two numbers: $2 + 3 = 5$), the AoRV describes rules for adding, subtracting, multiplying, and dividing random variables, based on distributional assumption (Springer 1979). This approach also allows to perform other commonly used aggregation function, such as calculating averages, minimum, and maximum values of sets of several random variables.

First, we introduce the general concepts, relying on the work of Whittle (2000). A random variable is a function that defines the occurrence of all possible outcomes of a random phenomenon. Generally, a random phenomenon can be understood in terms of the probability distribution of possible outcomes, i.e., the function that gives the likelihood of a particular outcome (for discreet random phenomena) or a range of outcomes (for continuous random phenomena). The probability distribution may consequently be regarded as a property of a random variable, assigning probabilities to each possible outcome.

The AoRV describes how calculations with multiple random variables are performed. While the AoRV is not restricted to specific probability distributions, it assumes, however, that the distribution that describes the behavior of the initial random variables (i.e., the variables that we start to calculate with) is known. We can then describe the new random variable that results from the mathematical operations in terms of its probability distribution. In the following, for the sake of simplicity, we only describe the operations for statistically independent random variables. In general, all manipulations in the AoRV

presented in this paper can be extended to statistically dependent variables if their co-variances are known (Springer 1979).

First, adding two random variables results in a random variable that is described by the convolution of the original probability distributions. A convolution is a mathematical procedure that creates a new function out of two continuous functions f and g. Given two independent random variables X and Y, the convolution of the respective probability distribution functions (e.g., f and g) results in the probability density function describing the random variable Z = X+Y. This function is defined as

$$(f * g)(x) = \int_{-\infty}^{\infty} f(\tau)g(x-\tau)d\tau = \int_{-\infty}^{\infty} f(x-\tau)g(\tau)d\tau.$$

Multiplying two random variables will result in a new random variable that follows the product distribution, resulting from the product of the respective distribution functions of the original random variables. The probability density function of a product distribution, describing random variable Z as the product of two random variables X and Y (with their respective density functions), is defined by

$$f_z(z) = \int_{-\infty}^{\infty} f_X(x) f_Y(z/x) \frac{1}{|x|} dx.$$

Similarly, the ratio distribution defines the probability distribution resulting from the division of two random variables X and Y, and their respective distributions. This ratio distribution is calculated as

$$f_z(z) = \int_{-\infty}^{\infty} |y| f(zy, y) dx.$$

In the case of addition, the sum of two normally distributed random variables (i.e., the convolution of their distributions) is again normally distributed. This is, however, not the case for the product distribution that results from multiplying two normally distributed random variables. Consequently, more complex calculations in the AoRV can often only be described through lengthy and complex integrals that describe the probability distribution in terms of its density function. This makes calculations in the AoRV sometimes difficult to handle for humans, but computer programs can process and evaluate such complex integrals. To facilitate our analysis and case evaluation, we developed a prototype application with R (version 3.3.3., relying on the distr-package version 2.6.2), that allows us to perform calculations with random variables and the underlying distributions according to the AoRV.

3 Case Evaluation

In this section, we present and discuss two typical cases to evaluate the applicability and practical utility of applying the AoRV to calculations in real-world business analytics scenarios with unreliable data. Both cases were developed in close collaboration with experts of the field who provided the data used for the calculations, additional information on how this data is acquired, and how the calculation results are utilized. In the following, we changed some of the initial data input in the paper at hand to ensure anonymity. However, the actual scenarios and aggregation logics correspond to real-world scenarios that have a major impact on decisions within the respective companies. All calculations as well as the Monte Carlo simulations were implemented in R (Kohlas 1972).

The procedure for both cases was as follows: First, we collected data on a specific use case which is related to major financial decisions in the company. Then, to estimate the required distributional assumptions for our calculations, we asked experts to provide realistic worst-case scenarios for all input data (i.e., data that is not a consequence of any calculations or aggregations) and to estimate the likelihood of these scenarios. We then calculated matching distributions that were centered around the initial (assumed) precise values and fit the estimated likelihood of the lower and upper scenarios. Subsequently, we performed a series of calculations in which we replaced the numerical estimates that were used in real-world business analytics systems with distributional estimates based on the AoRV.

Furthermore, we used these estimated negative effects to develop a simple Monte-Carlo simulation, that demonstrates the potential financial benefit that could come from a more adequate treatment of unreliability in the two discussed business analytics cases. After our calculations, we again consulted the experts to estimate the negative financial impact of deviating from the expectations. This was done by

discussing the five worst scenarios from 1,000 simulation runs, in combination with the overall statistical estimates derived by applying the AoRV.

3.1 Estimating the Effect of Employee Retention

In our first case, we estimate the effects of changes in employee retention rates on operating profit for a major European software company, in the following called SoftComp. We chose this case because it nicely illustrates our preceding argument that adequately treating unreliability is particularly important for business analytics scenarios that involve non-financial and non-material assets, since data inputs are frequently only rough estimates (Rindova et al. 2010).

Senior management in SoftComp wants to evaluate the direct (in terms of HR costs and vacancy as well as integration effects) financial impact of potential organizational changes that are expected to decrease the employee retention rate. To that end, the controlling department collected internal data (modified in this manuscript to be the average of multiple companies) that estimates

1. the expected change in the employee retention rate (~1%),
2. the average recruitment costs per employee (~7,984 €),
3. the average costs incurred due to integration efforts per employee (~19,990 €),
4. the average labor costs saved per employee (~20,305 €),
5. the lost revenue from uncompleted labor due to vacancies per employee (~33,790 €),
6. the lost revenue from uncompleted labor due to integration efforts per employee (~27,199 €).

Based on adjusted forecasts of business operations, SoftComp then estimates the expected number of employees that leave the company. Subsequently, the overall effects can be calculated by multiplying the number of employees with the estimated averages, adding the resultant costs and revenue effects, and finally summing everything up. *Figure 1* displays this calculation; bold numbers in the middle of each cell show the initial estimates and the symbols on the lines indicate the performed operations.

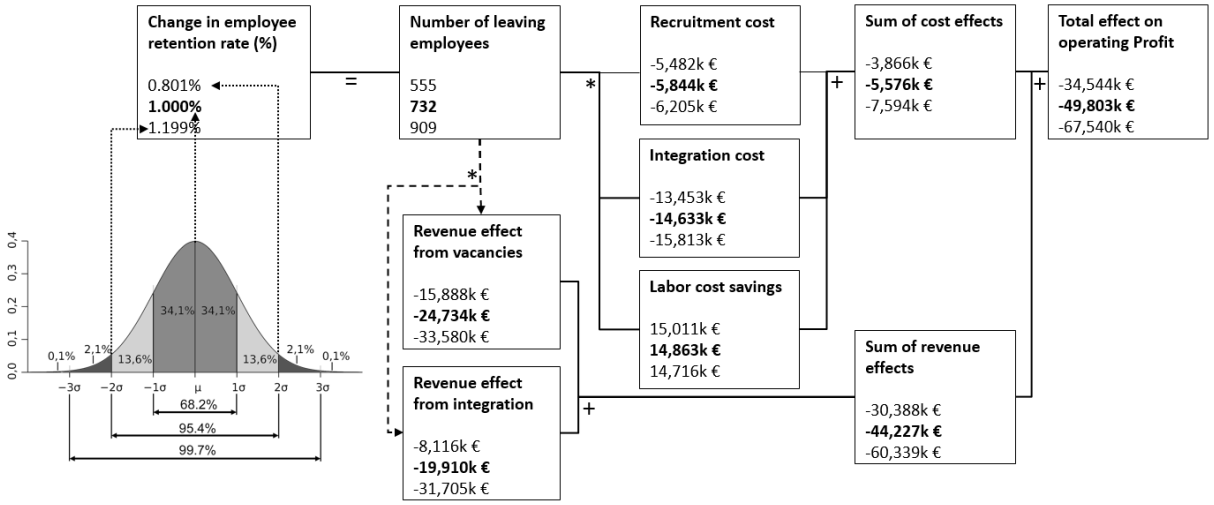

Figure 1. Estimating the effect of changes in employee retention rate on operating profit.

We interviewed several experts who were involved in the creation of these calculations about the reliability of the initial estimates. For some data points, such as the recruitment costs, integration costs, and labor costs, they believe that SoftComp is quite accurately tracking average values. Furthermore, SoftComp undertook a significant effort to validate the estimated effect of 1% employee retention, and thus expects only minor deviations. Regarding revenue effects, however, experts clearly stated that these numbers are rough estimates. Based on estimated worst-case scenarios and likelihoods, we then fit corresponding normal distributions to all initial data points, with the lower and upper 2σ-intervalls (~95% confidence) displayed in *Figure 1* surrounding the bold initial estimates. For example, from *Figure 1* we can see that there is a 95% chance that the overall number of leaving employees will be between 555 (best case, always above the bold number) and 909 (worst-case, always below the bold number).

Subsequently, we repeated the original calculations using random variables and the derived distributional assumptions. In Figure 1 we thus also display the 2σ-intervalls (~95% confidence) for all resultant constructs. Note, for example, that adding two random variables is a convolution of the underlying distributions (see Section 3), which is not the same as adding the best and worst-case scenarios. As key result, we obtained that the total effect on the operating profit is expected to vary between -34,544k € and -67,540k € with 95% confidence, which is quite a significant variance.

Table 1 below displays the five worst cases from 1,000 simulation runs. This analysis demonstrates an important point about complex models: to have catastrophic effects (~80 mio € less operating profit versus the expected ~50 mio €) not all things need to go wrong; it is enough, if several key assumptions, which interact accordingly, are wrong. Consider, for example, the first row in *Table 1*. Even though labor savings are quite high and integration effects are comparatively moderate, a high number of leaving employees in combination with significantly underestimated vacancy effects would lead to very high overall negative effects. In contrast, row 3 in *Table 1* shows a very different scenario, which has a moderate number of leaving employees and vacancy effects, but causes similarly bad overall effects due to high integration effects.

Employees leaving (#)	Recruitment costs	Integration costs	Labor savings	Vacancy effects	Integration effects	Cost effects	Revenue effects	Overall effects
949	-7,709	-18,623	19,214	-43,534	-31,493	-7,118	-75,027	-82,145
905	-7,042	-20,223	18,351	-37,425	-35,752	-8,914	-73,177	-82,091
847	-7,111	-16,843	17,222	-34,739	-39,813	-6,732	-74,552	-81,284
932	-6,991	-19,422	18,901	-36,496	-35,117	-7,512	-71,613	-79,125
860	-6,808	-17,384	17,409	-37,524	-33,290	-6,783	-70,814	-77,597

Table 1. Overview of worst cases from simulation. All values in 1,000 € (except employees).

3.2 Predictive Product Costing

The second case concerns a major European manufacturing company (in the following referred to as ManComp) that is required to estimate the overall costs that manufacturing a new product will incur. Quick and accurate costing estimates are key for the survival of the company, since margins are low and partner companies expect quick responses to manufacturing inquiries. For such estimates, ManComp currently uses a costing system that contains all (sub-)components bought from suppliers, all related labor processes, and depreciation and operation costs of machines. The system has been developed as an integrated software tool, collecting data automatically from internal databases. Each item (components, labor process, and depreciation) involved in production is assigned a corresponding unit price. Labor and machine time, for example, are commonly measured in costs per minute, whereas most other components are tracked as unit or mass prices. All related components are then aggregated (or "rolled-up") by the costing system, resulting in the total cost that is incurred in manufacturing the final product. The system also tracks additional overhead costs associated with sales, marketing, and other activities related to a specific product.

To evaluate the applicability of the AoRV in this case, we re-analyzed the cost of an order that the company received, requesting 10 units of the pump P-100 produced by ManComp. The product consists of three major components (Casing, Drive, and Shaft), that are assembled from a variety of subcomponents; see *Table 2* for the descriptions of all components, and corresponding quantities (Qty), units of measure (UoM) and unit and total costs. The pump P-100 is a moderately complex product, so that we only list the sub-components, work stages, and machine efforts of the casing in *Table 2* due to space constraints. The price for components and the duration of labor processes are based on historical data that is tracked in other related business analytics systems. The depreciations of a machines' values are based on accounting assumptions about the machines' lifetime and erosion.

Description	Qty	UoM	Cost per Unit [€]	Total Cost [€]	95% Confidence (high)	95% Confidence (low)
Pump P-100	10	PC	811.50	**8'115.02**	**8'512.65**	**7'657.39**
Overhead estimate	1		73.77	737.73	939.14	536.32
Work Center sales processing	100	min	3.60	360.00	420.42	299.58
Casing	10	PC	455.12	4'551.20	4'882.50	4'159.90
Slug for casing	10	PC	11.00	110.00	120.07	99.93
Turn casing (setup)	10	min	2.40	24.00	34.07	13.93
Turn casing (labor)	300	min	3.60	1'080.00	1'382.11	777.89
Turn casing (machine)	300	min	3.60	1'080.00	1'180.71	979.30
Drill holes (setup)	3	min	2.40	7.20	11.23	3.17
Drill holes (labor)	100	min	3.60	360.00	480.85	239.15
Drill holes (machine)	100	min	3.60	360.00	420.42	299.58
Flat seal	10	PC	11.00	110.00	120.07	89.58
Insert flat seal (labor)	50	min	3.60	180.00	210.42	89.58
Hexagon screw	80	PC	11.00	880.00	900.14	859.86
Inspect and deliver (labor)	100	min	3.60	360.00	440.56	279.44
Drive	10	PC	225.04	2'250.40	2'330.96	2'169.84
…	…	…	…	…	…	…
Shaft	10	PC	21.57	215.69	255.97	174.41
…	…	…	…	…	…	…

Table 2. Cost positions and calculation of ManComp`s pump P-100

To gain some understanding of the reliability of their data, ManComp tracks "confidence levels" for the basic components (everything not in italics in *Table 2*). These confidence levels are numbers between 1 (low confidence) and 5 (high confidence) that are manually assigned based on experience. We employed this data to grade the reliability of the estimated cost of specific items. Note that this estimation is currently only performed for the basic components (i.e., not for the aggregated costs). This is, however, unsatisfactory, as one expert, who is responsible for the design of the costing system, points out in his description of the system:

"*Every part and material has a price. The time for every step, for every activity, is documented. All resources, that interact when producing a part, are listed; machines as well as labor. The component lists and work plans are either from the ERP system or entered manually. Estimates are very different: some prices come from previous experiences, others are only very roughly estimated, so there is a lot of insecurity involved. We want to track this. At the end of the day, we do not only need to know that, overall, this will cost 10.000 €, but that with a 95% confidence we will be between 10.130 € and 10.624 €. That is what we really want – some sort of confidence interval."*

Consequently, we again calculated a 95% confidence interval for all basic values, estimated by the original confidence levels (see the last two columns of *Table 2)*. We then employed calculations based on the AoRV to estimate confidence intervals for the aggregated components, and finally, for the overall cost of producing ten P-100 pumps (see the bold numbers in *Table 2*). Furthermore, we conducted 1,000 runs of a Monte Carlo simulation, to better understand how insecurities in the basic data inputs interact during the aggregation. The results are quite different from the SoftComp scenario: since we mostly conducted additions and no multiplications, the variations in the single positions tend to cancel each other out rather than to lead to compounding effects. This is not surprising, but rather illustrates the comparatively low variance in the aggregated estimated cost of the pump in *Table 2*. Overall, the lowest estimate in our 1,000 simulation runs was a price of 7,446.13€ and the highest estimate was 8,707.30€.

4 Discussion and Outlook

We applied the AoRV in two different business analytics scenarios to enhance previously conducted analyses with additional capability to track the impact of unreliable data inputs on calculated results. This demonstrates that it is indeed possible to apply the theoretical foundations provided by the AoRV

to real-world business analytics problems to solve common problems encountered in practice. In fact, we can provide exactly the type of confidence interval analysis that was requested by the product costing expert in the ManComp case (see the quote in Section 4.2).

In both cases, we discussed the added value of our results with subject matter experts. In the first case, the application of the AoRV provides SoftComp with a better understanding about how extreme scenarios might look like, and what they would need to be prepared for, should some of their initial estimates be wrong. Here, we could observe that variations in the input variables did not cancel each other out, but rather showed compounding effects. Consequently, the estimated worst-case (~ 68 mio €) is almost twice as costly as the estimated best case (~ 35 mio €) of the 95% confidence interval (see *Figure 1*). Furthermore, the Monte Carlo simulation provided illustrative scenarios for such worst cases, highlighting how different combinations of false estimates can lead to similarly bad results.

In contrast, applying the AoRV to the product cost estimation scenario of ManComp demonstrates that this approach can be used to estimate potential cost ranges and risks more precisely. As one expert stated: "*[We] receive a request and need to produce an offer within a very short time. This offer needs to be competitive, but at the same time, we must not make a loss with it. Here, I need to know how reliable the overall initial estimate is.*" Due to the comparatively low margins of ManComp, precise estimates and risk analyses are of crucial importance for the survival of the company. Experts also mentioned an interesting future direction for this research, which could aim at employing the AoRV approach to identify the initial data points that have the overall highest impact on the calculated overall reliability. Since the time for producing price offers is limited and since improving data quality is both costly and time consuming, a more efficient and targeted approach to data quality management would be of high value for ManComp.

Overall, we find that it is difficult to predict a-priori how uncertainties interact in complex business analytics scenarios, whether they produce potentially dangerous compounding effects (e.g., the SoftComp scenario) or whether the different uncertainties tend to cancel each other out, leading to relatively stable and predictable overall effects (e.g., the ManComp scenario). The AoRV does not directly improve the reliability of input data. Instead, calculations with the AoRV provide an algorithmic approach to mathematically asses data reliability in such scenarios. This provides valuable design knowledge for business analytics applications (Baskerville et al. 2015).

We note two crucial limitations of the AoRV approach. First, it requires to estimate probability distributions for all initial data. This is not commonly known or tracked in current business analytics systems. Consequently, our work should only be considered a complementary effort to other research endeavors that are currently taking place in machine learning. Here, several scholars aim to understand how algorithms can be employed to discover regularities in larger datasets. These machine learning approaches allow to estimate probability distributions, for example, based on historical data. For developing practically useful automated applications, such techniques could be combined with the AoRV technique. Second, our approach is limited to numerical (or at least ordinal) data, for which the mathematical operations make sense. Consequently, it does not apply to reliability issues that occur, for example, in aggregating categorical customer location data (such as countries, cities, street names).

As a next step of this design science research project (Peffers et al. 2007), we plan to develop and test an AoRV-based extension to a commonly used business analytics application. Currently, by retrospectively employing AoRV to solve past problems, we risk unconsciously modifying our assumptions to fit the expected outcomes, thereby threatening the validity of our evaluation (Barratt et al. 2011). Thus, we plan the next iteration of this research to involve real users, real data, and real current problems (Venable et al. 2012).

References

Abraham, R., Aier, S., and Winter, R. 2014. "Fail Early, Fail Often: Towards Coherent Feedback Loops in Design Science Research Evaluation," in *Proceedings of the International Conference on Information Systems - Building a Better World through Information Systems*, AIS Electronic Library (AISeL): Association for Information Sytems, December 14. (http://aisel.aisnet.org/icis2014/proceedings/ISDesign/3/).

Ballou, D. P., and Pazer, H. L. 1995. "Designing Information Systems to Optimize the Accuracy-Timeliness Tradeoff," *Information Systems Research* (6:1), pp. 51–72. (https://doi.org/10.1287/isre.6.1.51).

Barratt, M., Choi, T. Y., and Li, M. 2011. "Qualitative Case Studies in Operations Management: Trends, Research Outcomes, and Future Research Implications," *Journal of Operations Management* (29:4), pp. 329–342. (https://doi.org/10.1016/j.jom.2010.06.002).

Baskerville, R. L., Kaul, M., and Storey, V. C. 2015. "Genres of Inquiry in Design-Science Research: Justification and Evaluation of Knowledge Production," *MIS Quarterly* (39:3), pp. 541-A9.

Bose, R. 2009. "Advanced Analytics: Opportunities and Challenges," *Industrial Management & Data Systems* (109:2), pp. 155–172. (https://doi.org/10.1108/02635570910930073).

Davenport, T. H. 2006. "Competing on Analytics," *Harvard Business Review* (84:1), pp. 1–10.

DeLone, W. H., and McLean, E. R. 1992. "Information Systems Success: The Quest for the Dependent Variable," *Information Systems Research* (3:1), pp. 60–95.

Draper, D. 1995. "Assessment and Propagation of Model Uncertainty," *Journal of the Royal Statistical Society. Series B (Methodological)* (57:1), pp. 45–97.

Eisenhardt, K. M., and Zbaracki, M. J. 1992. "Strategic Decision Making," *Strategic Management Journal* (13), pp. 17–37.

Elbashir, M. Z., Collier, P. A., and Davern, M. J. 2008. "Measuring the Effects of Business Intelligence Systems: The Relationship between Business Process and Organizational Performance," *International Journal of Accounting Information Systems* (9:3), Eighth International Research Symposium on Accounting Information Systems (IRSAIS), pp. 135–153. (https://doi.org/10.1016/j.accinf.2008.03.001).

Embrechts, P., Puccetti, G., and Rüschendorf, L. 2013. "Model Uncertainty and VaR Aggregation," *Journal of Banking & Finance* (37:8), pp. 2750–2764.

Ford, C. M., and Gioia, D. A. 2000. "Factors Influencing Creativity in the Domain of Managerial Decision Making," *Journal of Management* (26:4), pp. 705–732.

Hernández, M. A., and Stolfo, S. J. 1998. "Real-World Data Is Dirty: Data Cleansing and the Merge/Purge Problem," *Data Mining and Knowledge Discovery* (2:1), pp. 9–37.

Jarke, M., Jeusfeld, M. A., Quix, C., and Vassiliadis, P. 1999. "Architecture and Quality in Data Warehouses: An Extended Repository Approach," *Information Systems* (24:3), 10th International Conference on Advanced Information Systems Engineering, pp. 229–253. (https://doi.org/10.1016/S0306-4379(99)00017-4).

Jarke, M., and Vassiliou, Y. 1997. "Data Warehouse Quality: A Review of the DWQ Project.," in *IQ*, pp. 299–313.

Kahn, B. K., Strong, D. M., and Wang, R. Y. 2002. "Information Quality Benchmarks: Product and Service Performance," *Communications of the ACM* (45:4), pp. 184–192.

Kohlas, J. 1972. *Monte Carlo Simulation im Operations Research*, Berlin, New York: Springer.

Lee, Y. W., Strong, D. M., Kahn, B. K., and Wang, R. Y. 2002. *AIMQ: A Methodology for Information Quality Assessment*.

Madnick, S. E., Wang, R. Y., Lee, Y. W., and Zhu, H. 2009. "Overview and Framework for Data and Information Quality Research," *Journal of Data and Information Quality (JDIQ)* (1:1), p. 2.

Molloy, J. C., Chadwick, C., Ployhart, R. E., and Golden, S. J. 2011. "Making Intangibles 'Tangible' in Tests of Resource-Based Theory: A Multidisciplinary Construct Validation Approach," *Journal of Management* (37:5), pp. 1496–1518. (https://doi.org/10.1177/0149206310394185).

Murphy, K. R., and Davidshofer, C. O. 2001. *Psychological Testing: Principles and Applications*, Prentice Hall.

Negash, S., and Gray, P. 2008. "Business Intelligence," in *Business Intelligence*, International Handbooks Information System. (https://doi.org/10.1007/978-3-540-48716-6_9).

Parssian, A., Sarkar, S., and Jacob, V. S. 2004. "Assessing Data Quality for Information Products: Impact of Selection, Projection, and Cartesian Product," *Management Science* (50:7), pp. 967–982. (https://doi.org/10.1287/mnsc.1040.0237).

Peffers, K., Tuunanen, T., Rothenberger, M., and Chatterjee, S. 2007. "A Design Science Research Methodology for Information Systems Research," *J. Manage. Inf. Syst.* (24:3), pp. 45–77. (https://doi.org/10.2753/MIS0742-1222240302).

Petter, S., DeLone, W., and McLean, E. R. 2013. "Information Systems Success: The Quest for the Independent Variables," *Journal of Management Information Systems* (29:4), pp. 7–62. (https://doi.org/10.2753/MIS0742-1222290401).

Redman, T. C. 1998. "The Impact of Poor Data Quality on the Typical Enterprise," *Commun. ACM* (41:2), pp. 79–82. (https://doi.org/10.1145/269012.269025).

Rindova, V., Williamson, I., and Petkova, A. 2010. "Reputation as an Intangible Asset: Reflections on Theory and Methods in Two Empirical Studies of Business School Reputations," *Journal of Management* (36:3), pp. 610–619. (https://doi.org/10.1177/0149206309343208).

Shankaranarayan, G., Ziad, M., and Wang, R. Y. 2003. "Managing Data Quality in Dynamic Decision Environments: An Information Product Approach," *Journal of Database Management (JDM)* (14:4), pp. 14–32. (https://doi.org/10.4018/jdm.2003100102).

Springer, M. D. 1979. *The Algebra of Random Variables*, Wiley.

Venable, J., Pries-Heje, J., and Baskerville, R. 2012. "A Comprehensive Framework for Evaluation in Design Science Research," in *Design Science Research in Information Systems. Advances in Theory and Practice*, Lecture Notes in Computer Science, K. Peffers, M. Rothenberger, and B. Kuechler (eds.), Springer Berlin Heidelberg, pp. 423–438. (http://link.springer.com/chapter/10.1007/978-3-642-29863-9_31).

Venable, J., Pries-Heje, J., and Baskerville, R. 2016. "FEDS: A Framework for Evaluation in Design Science Research," *European Journal of Information Systems* (25:1), pp. 77–89. (https://doi.org/10.1057/ejis.2014.36).

Wang, R. Y., Reddy, M. P., and Kon, H. B. 1995. "Toward Quality Data: An Attribute-Based Approach," *Decision Support Systems* (13:3), Information Technologies and Systems, pp. 349–372. (https://doi.org/10.1016/0167-9236(93)E0050-N).

Wang, R. Y., and Strong, D. M. 1996. "Beyond Accuracy: What Data Quality Means to Data Consumers," *Journal of Management Information Systems* (12:4), pp. 5–33.

Whittle, P. 2000. *Probability via Expectation*, (4th ed.), NY: Springer. .

Zmud, R. W. 1978. "An Empirical Investigation of the Dimensionality of the Concept of Information" *Decision Sciences* (9:2), pp. 187–195. (https://doi.org/10.1111/j.1540-5915.1978.tb01378.x).